Evaluating the Robustness of Watermark-based Detection of AI-generated Content

Neil Gong
Department of Electrical and Computer Engineering
Department of Computer Science (secondary appointment)
Duke University
10/16/2023

Ethical Concerns of AI-generated Content

- Harmful content
- Disinformation and propaganda campaigns
- Teaching and education

Content Moderation for Generative AI

- Preventing generation of harmful content: safety filters
  - Deployed by many GenAI services
  - Yang et al. “SneakyPrompt: Evaluating Robustness of Text-to-image Generative Models' Safety Filters”. *Arxiv*, 2023.

- Detecting AI-generated content
  - Watermark-based detection of AI-generated images is deployed by Google, Stability AI, OpenAI, etc.
  - Jiang et al. "Evading Watermark based Detection of AI-Generated Content". In *ACM Conference on Computer and Communications Security (CCS),* 2023.

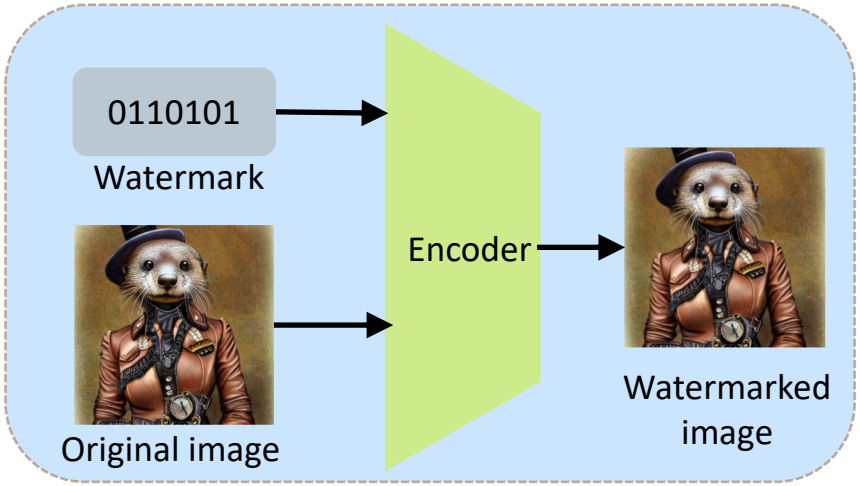
Detecting AI-generated Content

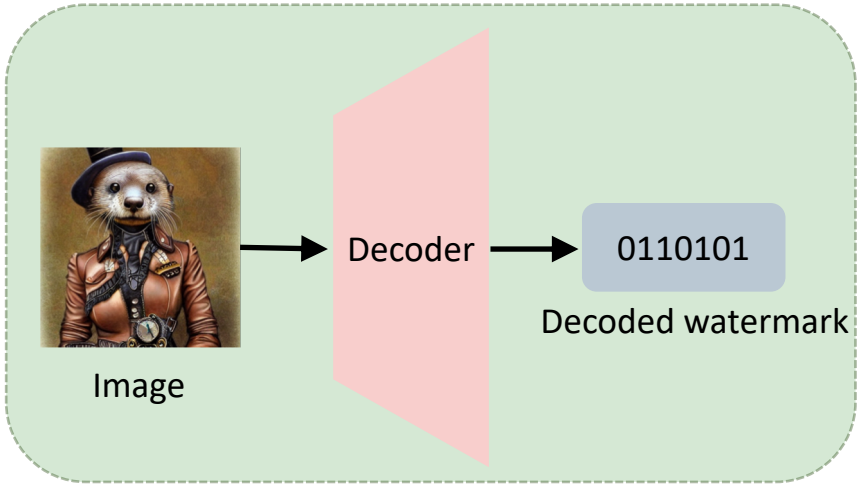
- Passive detection
  - Key idea: leverage artifacts in AI-generated content
  - High false positives/negatives

- Watermark-based detection
  - Multiple companies have deployed such detector
  - This talk
    - AI-generated images

Image Watermarking

- Three components
  - Watermark (bitstring)
  - Encoder
  - Decoder





Watermarking Methods

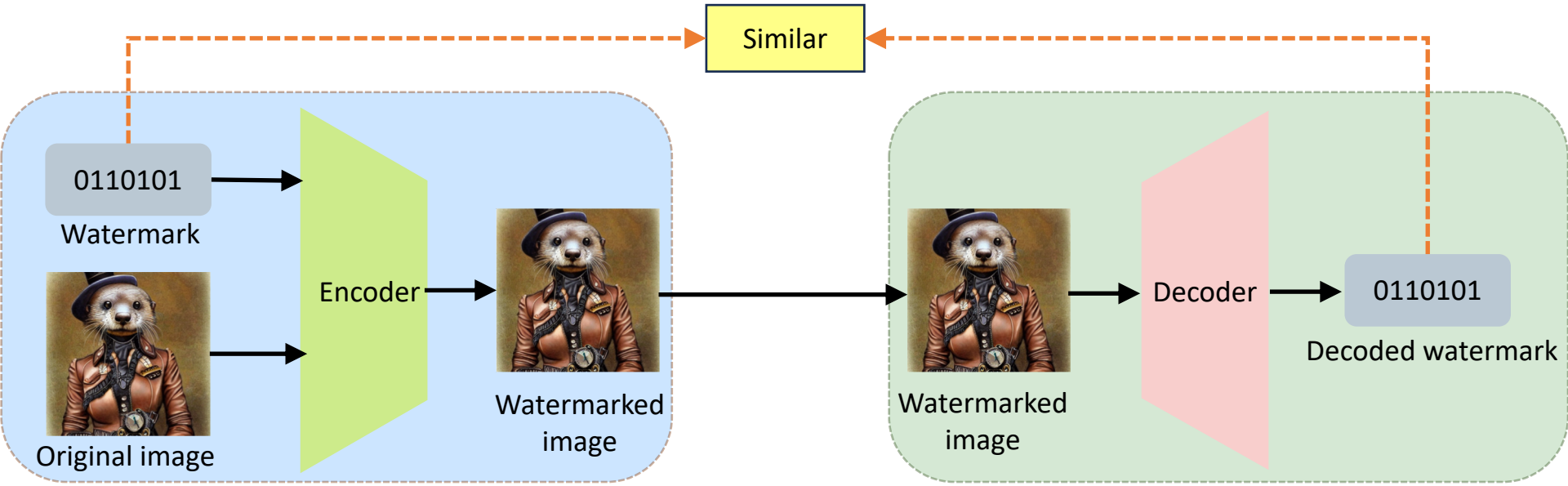
- Non-learning-based
  - Encoder and decoder are handcrafted based on heuristics
  - Not robust to common post-processing
    - JPEG compression, Gaussian noise, Gaussian blur, Brightness/Contrast

(a) Watermarked (b) JPEG

Watermark used by Stable Diffusion

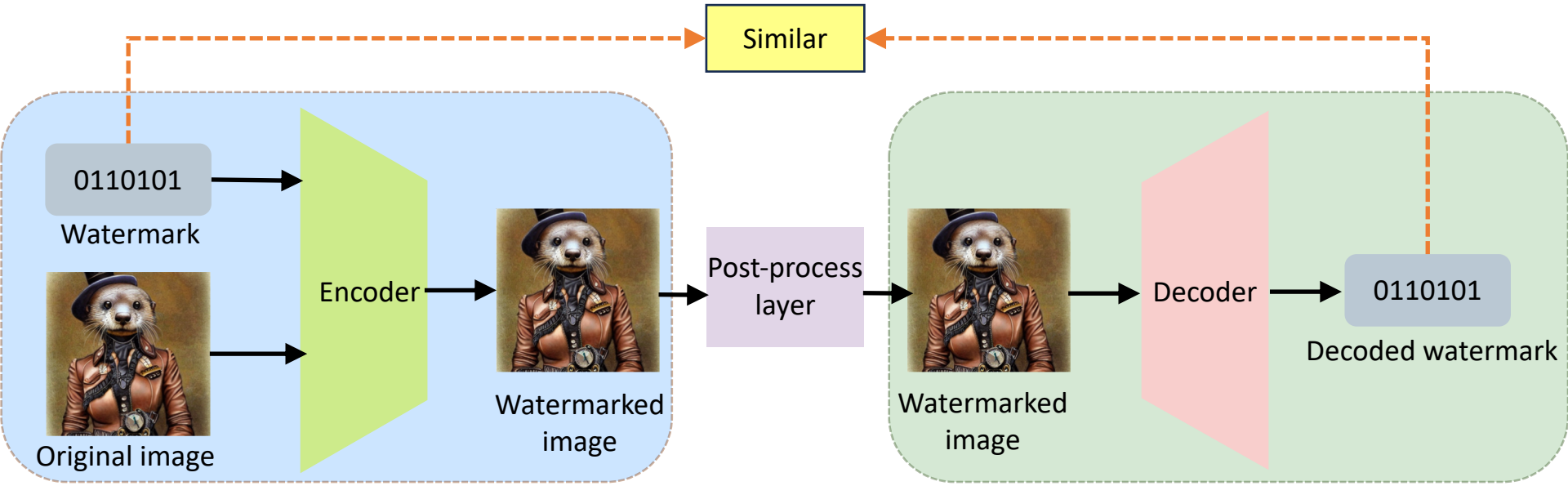
- Learning-based
  - Encoder and decoder are neural networks
  - Believed to be robust due to *adversarial training*

Standard vs. Adversarial Training



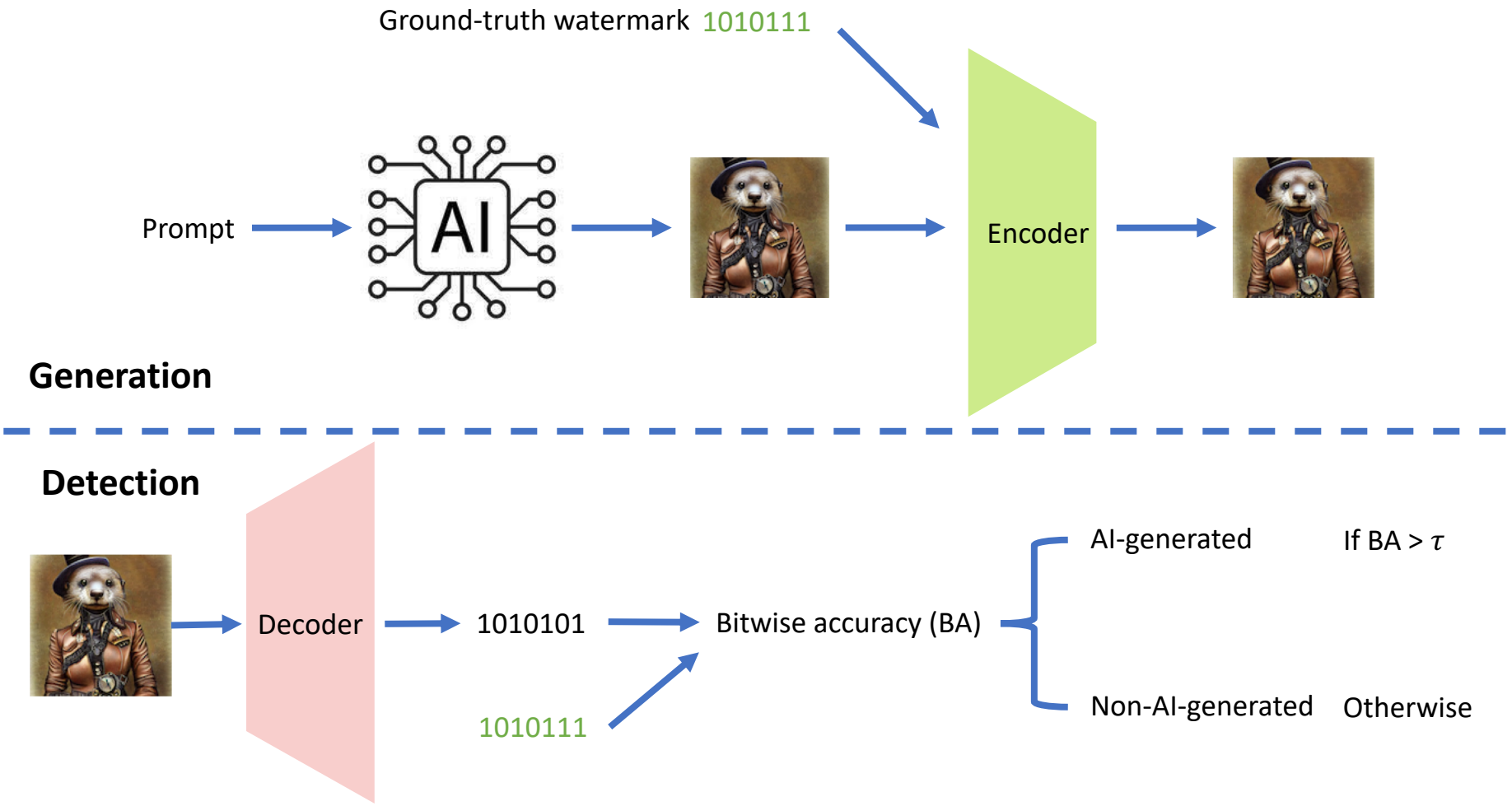
Standard training

Standard vs. Adversarial Training



Adversarial training

Watermark-based Detection



How to Set Detection Threshold τ?

- Achieve a desired False Positive Rate (FPR)

Non-AI-generated

Decoder

1010101

BA

1010111

Uniformly at random

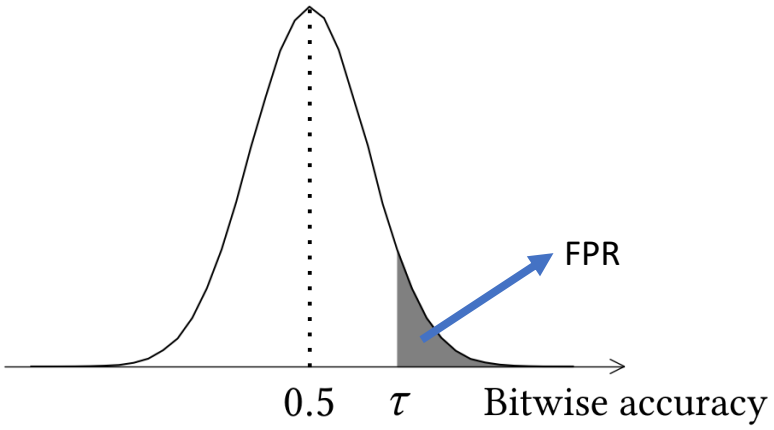
FPR < 10^{-4}

30-bit watermark

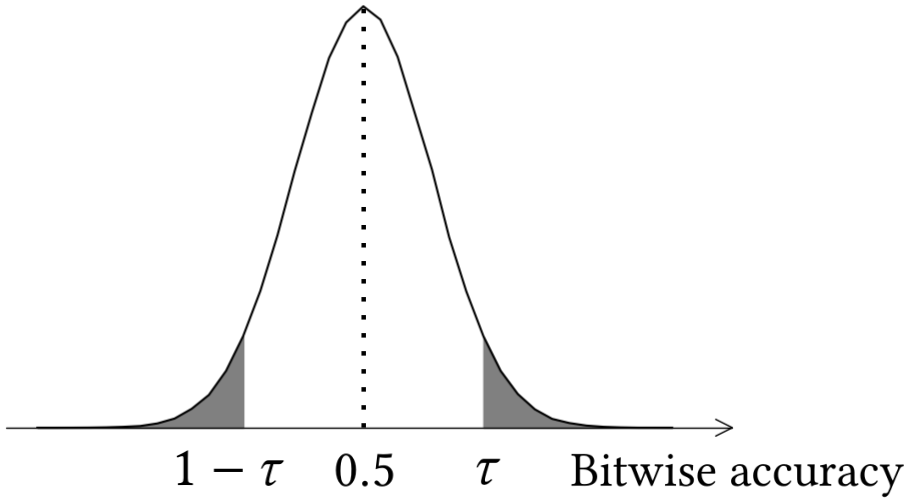
$\tau > 0.83$

n: watermark length

$n \cdot BA \sim Binomial\ (n, 0.5)$



Double-tail Detector



AI-generated	If BA > τ or BA < 1 - τ
Non-AI-generated	Otherwise

Detector Deployment Scenarios

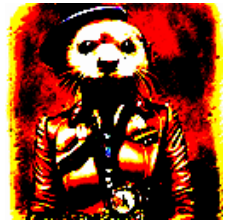
- Detection-as-a-service
  - Provider of GenAI service also provides detection service
- End-user detection
  - Detector as end-user app
    - Mobile app
    - Browser plugin
- Public detection
  - Publicly release decoder and watermark
  - Individuals can personalize τ depending on desired FPR
- Third-party detection
  - GenAI provider shares decoder and watermark with selected third parties
  - E.g., Google → Twitter

Threat Model

- White-box setting
  - Attacker has access to decoder
  - Aim to evade detector with any $\tau > 0.5$
- Black-box setting
  - Attacker has access to detector API
  - Aim to evade a specific detector with an unknown τ
- Focus on watermark removal
  - Watermark forging/spoofing is technically the same

One Visualization Example

Standard training

Original	Watermarked	JPEG	Gaussian noise	Gaussian blur	Brightness /Contrast	White box	Black box
							
							

Adversarial training

White-box Setting

- Given a watermarked image I_w
- Add minimal perturbation δ
- s.t. each bit of decoded watermark flips

$$\min_{\delta} ||\delta||_\infty$$

$$s.t.\ D(I_w + \delta) = \neg D(I_w)$$

Decoder

Flip each bit

Guaranteed to evade single-tail detector

Can be detected by double-tail detector

White-box Setting

- Intuition: non-watermarked images have bitwise accuracy ≈ 0.5
- Add minimal perturbation to make bitwise accuracy ≈ 0.5
  - Perturbed watermarked image indistinguishable with non-watermarked ones

$$\min_{\delta} ||\delta||_{\infty}$$

$$s.t.\ |BA(D(I_w + \delta), w) - 0.5| \leq \epsilon$$

Ground-truth watermark

White-box Setting

- Intuition: non-watermarked images have bitwise accuracy ≈ 0.5
- Add minimal perturbation to make bitwise accuracy ≈ 0.5
  - Perturbed watermarked image indistinguishable with non-watermarked ones

$$\min_{\delta} ||\delta||_{\infty}$$

$$s.t. \cancel{|BA(D(I_w + \delta), w) - 0.5| \leq \epsilon}$$

$$BA(D(I_w + \delta), w_t) \geq 1 - \epsilon$$

Ground-truth watermark

Random watermark
target watermark

Theoretical Evaluation

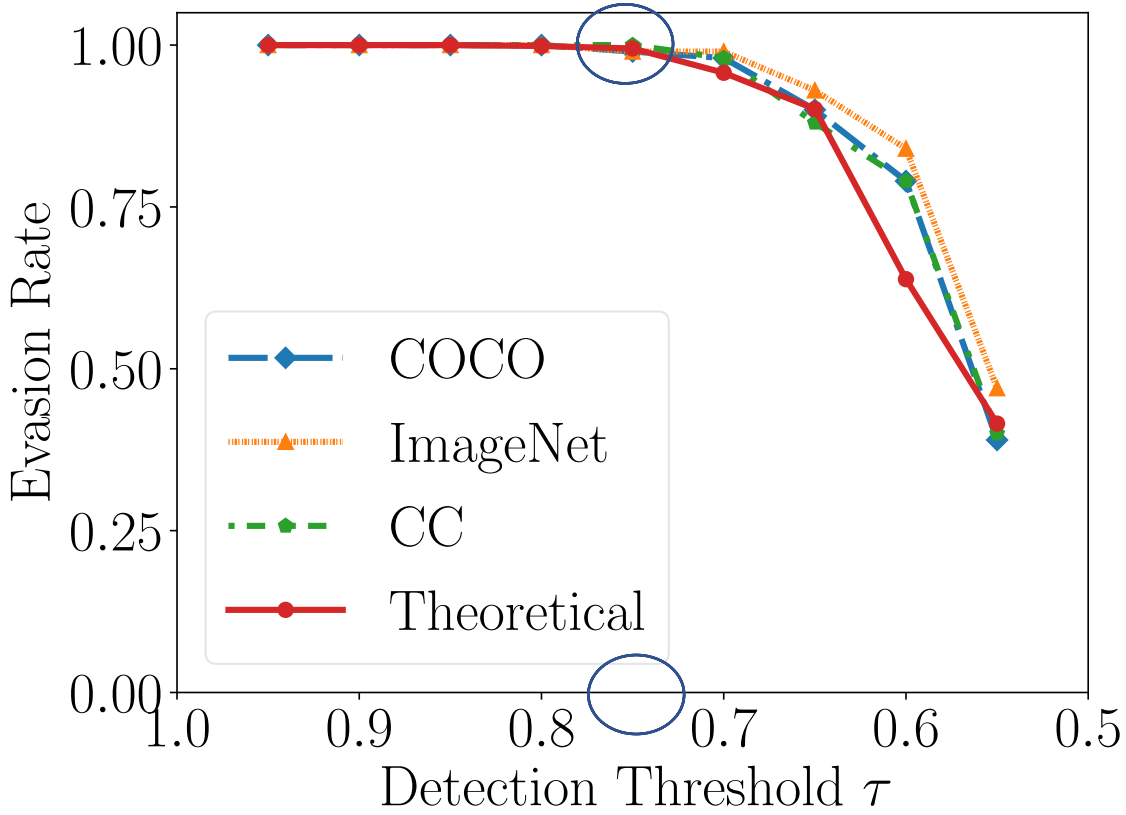
Evasion rate: probability that a perturbed watermarked image is detected as non-AI-generated

Lower bound for single-tail detector: $\Pr(m \leq \lfloor(\tau - \varepsilon)n\rfloor)$

$m \sim Binomial(n, 0.5)$

Lower bound for double-tail detector: $2\Pr(m \leq \lfloor(\tau - \varepsilon)n\rfloor) - 1$

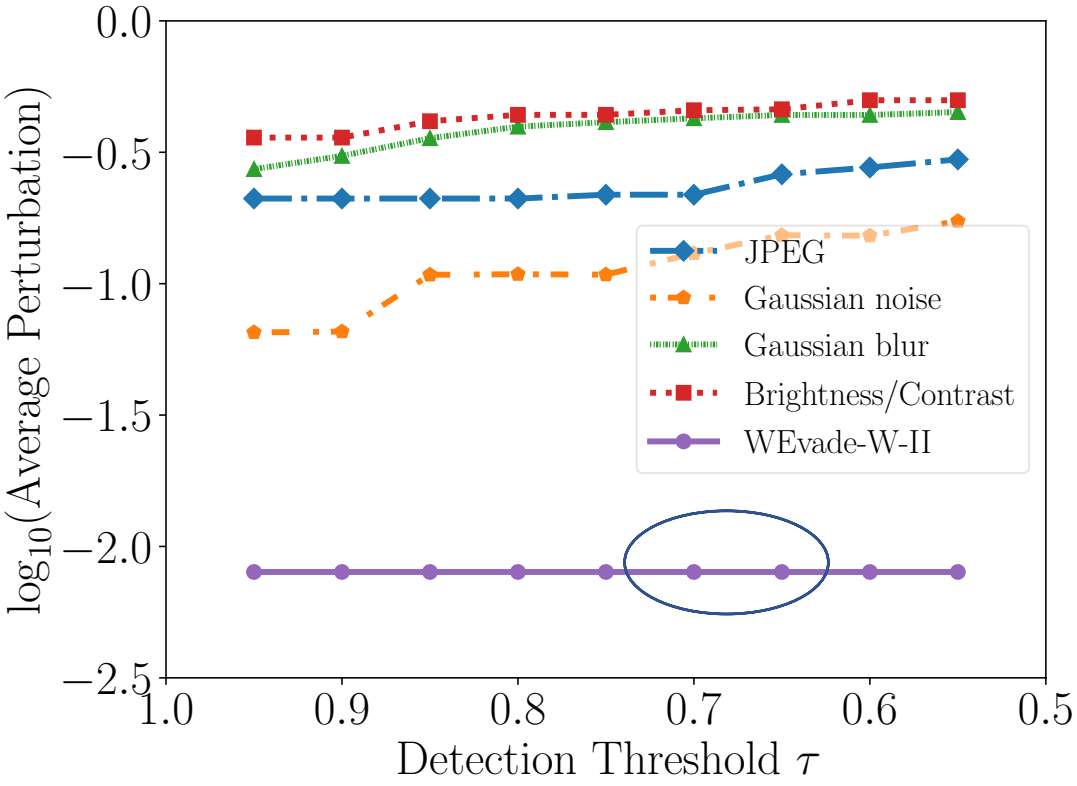
Empirical Evaluation Results



30-bit watermark, standard training

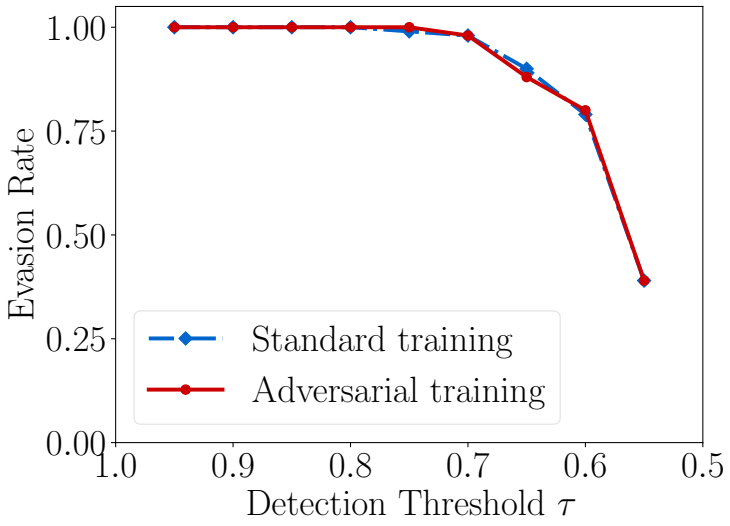
Double-tail detector

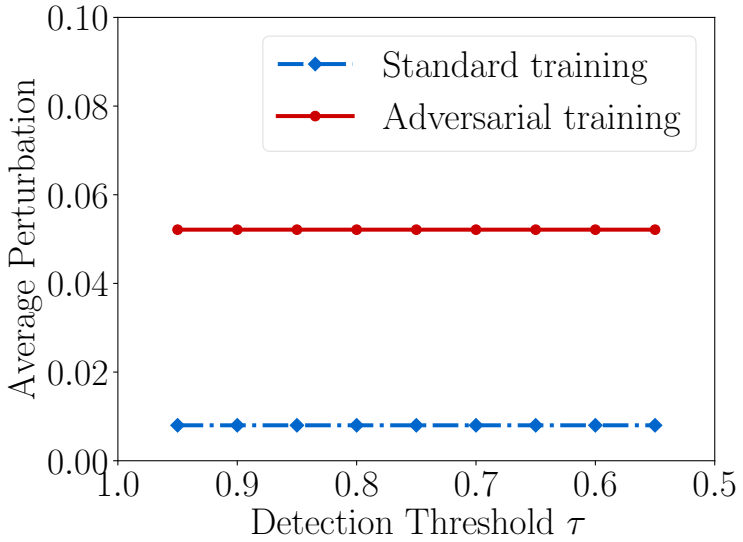
Our Adversarial Post-processing Adds Smaller Perturbations than Existing Ones



COCO dataset

Adversarial Training Improves Robustness

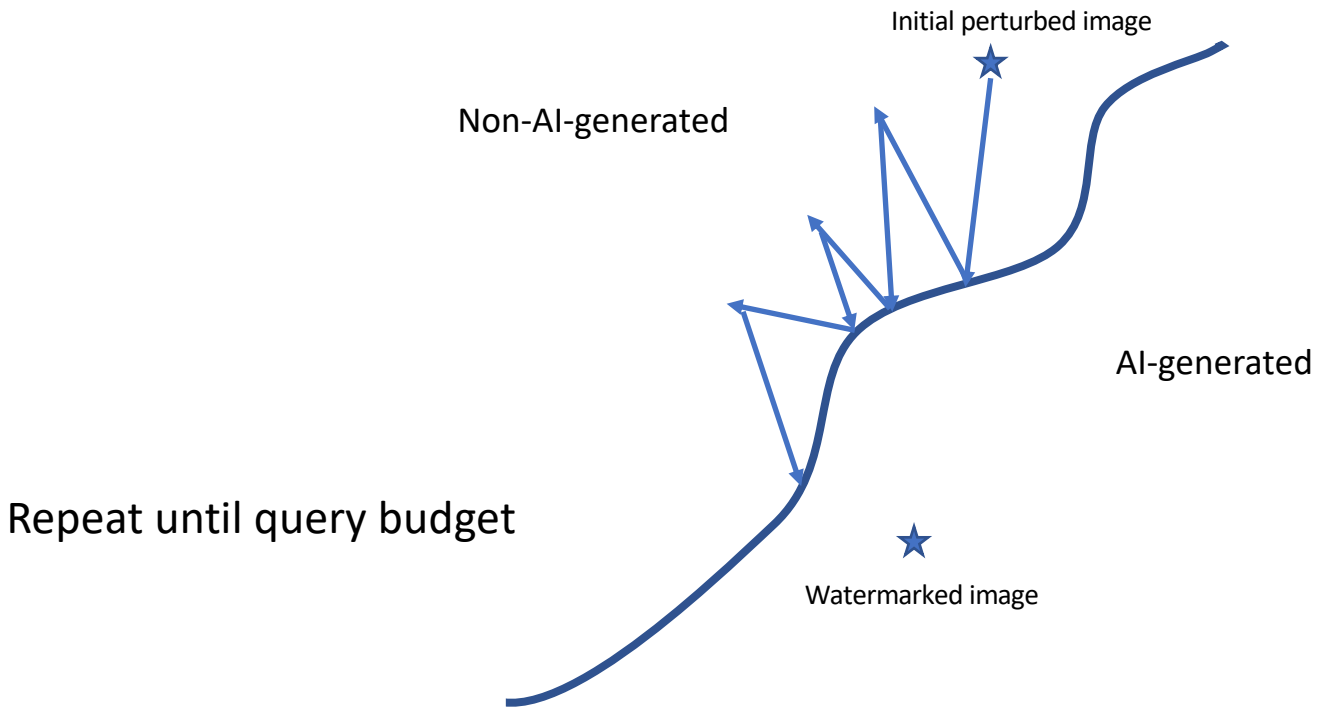




Take-away Messages

- Learning-based-watermark based detection has good robustness to common post-processing in *non-adversarial settings*
- Broken in the white-box setting in *adversarial settings*
- Adversarial training improves robustness but still insufficient

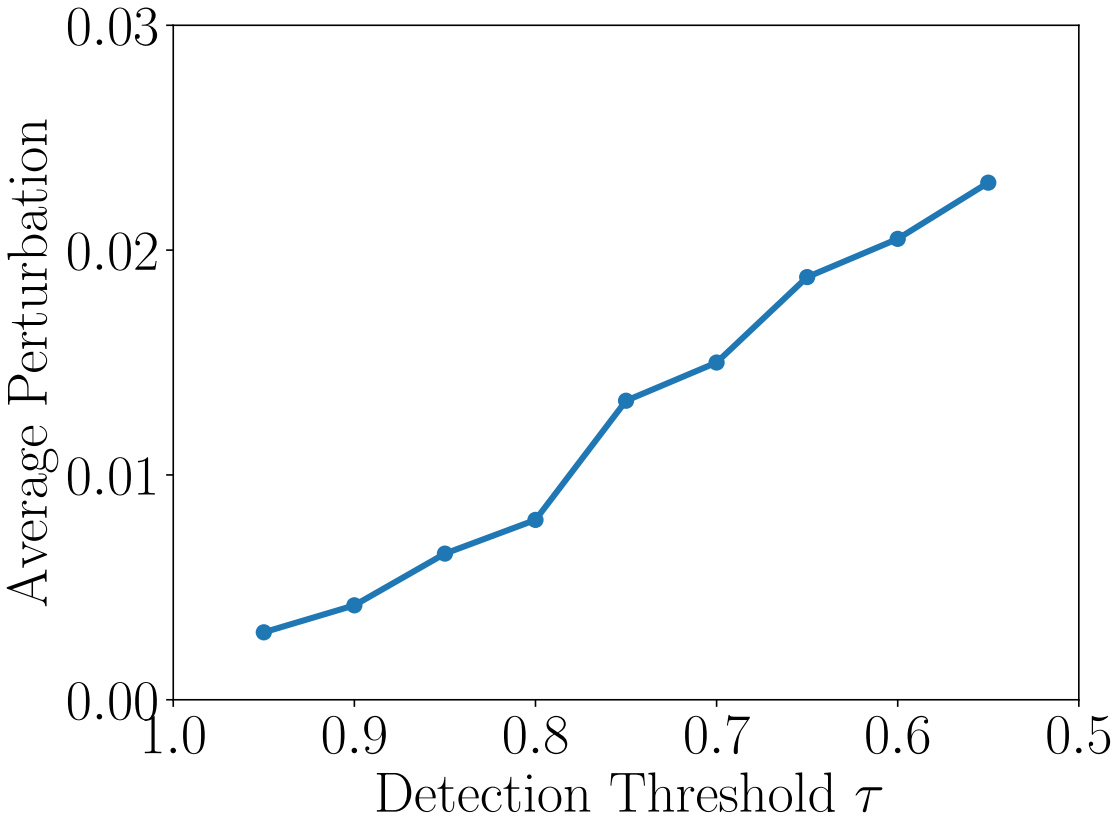
Black-box Setting



Theoretical Evaluation

Evasion rate is guaranteed to be 1

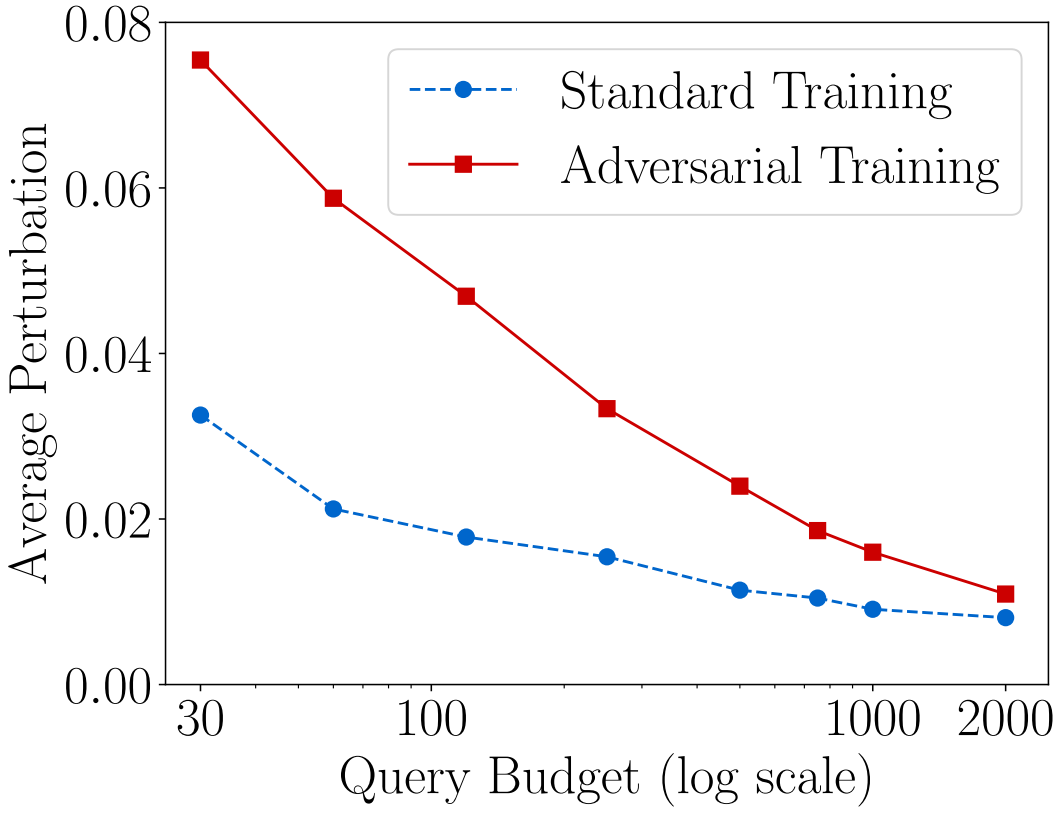
Empirical Evaluation Results



30-bit watermark, standard training, query budget 2,000

COCO dataset

Adversarial Training Improves Robustness



Summary and Discussion

- Don’t publicly release decoder
  - No white-box attack
- Adversarial training can improve robustness
- But probably still insufficient
  - Dozens-hundreds of queries to evade a black-box detector
  - While maintaining image quality
- Stronger adversarial training?
- Restricted access to detector API?
  - Attacker cannot access detector API
  - Transfer attack

Acknowledgements

Zhengyuan Jiang Jinghuai Zhang